## Amendment to the Claims

This listing of claims will replace all prior versions, and listings, of claims in the application:

## Listing of Claims:

1       1.      (currently amended): A system for grouping clusters of
2    semantically scored documents electronically stored in a data corpus, comprising:
3           a scoring module determining a score, which is assigned to at least one
4    concept that has been extracted from a plurality of electronically-stored
5    documents, wherein the score is ~~based on at least one~~ calculated as a function of a
6    summation of a frequency of occurrence of the at least one concept within at least
7    one such document, a concept weight, a structural weight, and a corpus weight,
8    forming the score assigned to the at least one concept as a normalized score vector
9    for each such document, and determining a similarity between the normalized
10   score vector for each such document as an inner product of each normalized score
11   vector;
12          a clustering module forming clusters of the documents, comprising:
13              a selection submodule ~~evaluating~~ selecting a set of candidate seed
14   documents ~~selected~~ from the plurality of documents;
15              a seed document identification submodule identifying a set of seed
16   documents by applying the similarity ~~as a best fit~~ to each such candidate seed
17   document and selecting those candidate seed documents that are sufficiently
18   unique from other candidate seed documents as the seed documents;
19              a non-seed document identification submodule identifying a
20   plurality of non-seed documents;
21              a comparison submodule determining the similarity between each
22   non-seed document and a center of each cluster; and
23              a clustering submodule grouping each such non-seed document
24   into a cluster with [[the]] a best fit, subject to a minimum fit;

25      a threshold module determining the similarity between each of the

26    documents grouped into each cluster based on the center of the cluster and the

27    scores assigned to each of the at least one concepts in that document, dynamically

28    determining a threshold for each cluster as a function of the similarity between

29    each of the documents, and identifying and reassigning each of the documents

30    having the similarity falling outside the threshold.

1      Claim 2 (canceled).

1      3.    (currently amended): A system according to ~~Claim 2~~ <u>Claim 1</u>,

2    further comprising:

3      a compression module compressing the score through logarithmic

4    compression.

1      4.    (original): A system according to Claim 1, further comprising:

2      the scoring module calculating the concept weight as a function of a

3    number of terms comprising the at least one concept.

1      5.    (original): A system according to Claim 1, further comprising:

2      the scoring module calculating the structural weight as a function of a

3    location of the at least one concept within the at least one such document.

1      6.    (original): A system according to Claim 1, further comprising:

2      the scoring module calculating the corpus weight as a function of a

3    reference count of the at least one concept over the plurality of documents.

1      Claims 7-8 (canceled).

1      9.    (currently amended): A method for grouping clusters of

2    semantically scored documents electronically stored in a data corpus, comprising:

3      determining a score, which is assigned to at least one concept that has

4    been extracted from a plurality of electronically-stored documents, wherein the

5    score is ~~based on at least one~~ <u>calculated as a function of a summation</u> of a

6     frequency of occurrence of the at least one concept within at least one such

7     document, a concept weight, a structural weight, and a corpus weight;

8          forming the score assigned to the at least one concept as a normalized

9     score vector for each such document;

10          determining a similarity between the normalized score vector for each

11     such document as an inner product of each normalized score vector;

12          forming logically-grouped clusters of the documents, comprising:

13             ~~evaluating~~ <u>selecting</u> a set of candidate seed documents ~~selected~~

14     from the plurality of documents;

15             identifying a set of seed documents by applying the similarity ~~as a~~

16     ~~best fit~~ to each such candidate seed document <u>and selecting those candidate seed</u>

17     <u>documents that are sufficiently unique from other candidate seed documents as</u>

18     <u>the seed documents;</u>

19             identifying a plurality of non-seed documents;

20             determining the similarity between each non-seed document and a

21     center of each cluster; and

22             grouping each such non-seed document into a cluster with [[the]] <u>a</u>

23     best fit, subject to a minimum fit;

24          determining the similarity between each of the documents grouped into

25     each cluster based on the center of the cluster and the scores assigned to each of

26     the at least one concepts in that document;

27          dynamically determining a threshold for each cluster as a function of the

28     similarity between each of the documents; and

29          identifying and reassigning each of the documents having the similarity

30     falling outside the threshold.

1     Claim 10 (canceled).

1     11.    (currently amended): A method according to ~~Claim 10~~ <u>Claim 9</u>,

2     further comprising:

3     compressing the score through logarithmic compression.

1     12.     (original): A method according to Claim 9, further comprising:

2     calculating the concept weight as a function of a number of terms

3 comprising the at least one concept.

1     13.     (original): A method according to Claim 9, further comprising:

2     calculating the structural weight as a function of a location of the at least

3 one concept within the at least one such document.

1     14.     (original): A method according to Claim 9, further comprising:

2     calculating the corpus weight as a function of a reference count of the at

3 least one concept over the plurality of documents.

1     Claims 15-16 (canceled).

1     17.     (currently amended): A computer-readable storage medium

2 holding code for grouping clusters of semantically scored documents

3 electronically stored in a data corpus, comprising:

4     code for determining a score, which is assigned to at least one concept that

5 has been extracted from a plurality of electronically-stored documents, wherein

6 the score is ~~based on at least one~~ <u>calculated as a function of a summation</u> of a

7 frequency of occurrence of the at least one concept within at least one such

8 document, a concept weight, a structural weight, and a corpus weight;

9     code for forming the score assigned to the at least one concept as a

10 normalized score vector for each such document;

11     code for determining a similarity between the normalized score vector for

12 each such document as an inner product of each normalized score vector;

13     code for forming logically-grouped clusters of the documents, comprising:

14     code for ~~evaluating~~ <u>selecting</u> a set of candidate seed documents

15 ~~selected~~ from the plurality of documents;

16     code for identifying a set of seed documents by applying the

17 similarity ~~as a best fit~~ to each such candidate seed document <u>and selecting those</u>

18    candidate seed documents that are sufficiently unique from other candidate seed

19    documents as the seed documents;

20            code for identifying a plurality of non-seed documents;

21            code for determining the similarity between each non-seed

22    document and a center of each cluster; and

23            code for grouping each such non-seed document into a cluster with

24    [[the]] a best fit, subject to a minimum fit;

25            code for determining the similarity between each of the documents

26    grouped into each cluster based on the center of the cluster and the scores

27    assigned to each of the at least one concepts in that document;

28            code for dynamically determining a threshold for each cluster as a

29    function of the similarity between each of the documents; and

30            code for identifying and reassigning each of the documents having the

31    similarity falling outside the threshold.

1        18.    (currently amended): A system for providing efficient document

2    scoring of concepts within and clustering of documents in an electronically-stored

3    document set, comprising:

4            a scoring module scoring a document in an electronically-stored document

5    set, comprising:

6            a frequency module determining a frequency of occurrence of at

7    least one concept within a document;

8            a concept weight module analyzing a concept weight reflecting a

9    specificity of meaning for the at least one concept within the document;

10            a structural weight module analyzing a structural weight reflecting

11    a degree of significance based on structural location within the document for the

12    at least one concept;

13            a corpus weight module analyzing a corpus weight inversely

14    weighing a reference count of occurrences for the at least one concept within the

15    document;

16            a scoring evaluation module evaluating a score to be associated

17   with the at least one concept as a function of a summation of the frequency,

18   concept weight, structural weight, and corpus weight;

19            a vector module forming the score assigned to the at least one

20   concept as a normalized score vector for each such document in the

21   electronically-stored document set; and

22            a determination module determining a similarity between the

23   normalized score vector for each such document as an inner product of each

24   normalized score vector; .

25        a clustering module grouping the documents by the score into a plurality

26   of clusters, comprising:

27            a selection submodule ~~evaluating~~ selecting a set of candidate seed

28   documents ~~selected~~ from the electronically-stored document set;

29            a cluster seed submodule identifying seed documents by applying

30   the similarity ~~as a best fit~~ to each such candidate seed document and selecting

31   those candidate seed documents that are sufficiently unique from other candidate

32   seed documents as the seed documents;

33            an identification submodule identifying a plurality of non-seed

34   documents;

35            a comparison submodule determining the similarity between each

36   non-seed document and a cluster center of each cluster; and

37            a clustering submodule assigning each non-seed document to the

38   cluster with [[the]] a best fit, subject to a minimum fit; and

39        a threshold module relocating outlier documents, comprising determining

40   the similarity between each of the documents grouped into each cluster based on

41   the center of the cluster and the scores assigned to each of the at least one

42   concepts in that document, dynamically determining a threshold for each cluster

43   as a function of the similarity between each of the documents, and identifying and

44   reassigning each of the documents with the similarity falling outside the

45   threshold.

1        19.    (previously presented): A system according to Claim 18, further

2  comprising:

3        the scoring module evaluating the score in accordance with the formula:

4
$$S_i = \sum_{1 \to n}^{j} f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

5  where $S_i$ comprises the score, $f_{ij}$ comprises the frequency, $0 < cw_{ij} \leq 1$ comprises

6  the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$

7  comprises the corpus weight for occurrence $j$ of concept $i$.

1        20.    (previously presented): A system according to Claim 19, further

2  comprising:

3        the concept weight module evaluating the concept weight in accordance

4  with the formula:

5
$$cw_{ij} = \begin{cases} 0.25 + \left(0.25 \times t_{ij}\right), & 1 \leq t_{ij} \leq 3 \\ 0.25 + \left(0.25 \times \left[7 - t_{ij}\right]\right), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

6  where $cw_{ij}$ comprises the concept weight and $t_{ij}$ comprises a number of terms for

7  occurrence $j$ of each such concept $i$.

1        21.    (previously presented): A system according to Claim 19, further

2  comprising:

3        the structural weight module evaluating the structural weight in

4  accordance with the formula:

5
$$sw_{ij} = \begin{cases} 1.0, & if\left(j \approx SUBJECT\right) \\ 0.8, & if\left(j \approx HEADING\right) \\ 0.7, & if\left(j \approx SUMMARY\right) \\ 0.5 & if\left(j \approx BODY\right) \\ 0.1 & if\left(j \approx SIGNATURE\right) \end{cases}$$

6  where $sw_{ij}$ comprises the structural weight for occurrence $j$ of each such concept $i$.

1      22.    (previously presented): A system according to Claim 19, further

2  comprising:

3      the corpus weight module evaluating the corpus weight in accordance with

4  the formula:

5
$$rw_{ij} = \begin{cases} \left(\dfrac{T - r_{ij}}{T}\right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

6  where $rw_{ij}$ comprises the corpus weight, $r_{ij}$ comprises a reference count for

7  occurrence $j$ of each such concept $i$, $T$ comprises a total number of reference

8  counts of documents in the document set, and $M$ comprises a maximum reference

9  count of documents in the document set.

1      23.    (previously presented): A system according to Claim 19, further

2  comprising:

3      a compression module compressing the score in accordance with the

4  formula:

5      $S_i' = \log(S_i + 1)$

6  where $S_i'$ comprises the compressed score for each such concept $i$.

1      24.    (original): A system according to Claim 18, further comprising:

2      a global stop concept vector cache maintaining concepts and terms; and

3      a filtering module filtering selection of the at least one concept based on

4  the concepts and terms maintained in the global stop concept vector cache.

1      25.    (original): A system according to Claim 18, further comprising:

2      a parsing module identifying terms within at least one document in the

3  document set, and combining the identified terms into one or more of the

4  concepts.

1      26.    (original): A system according to Claim 25, further comprising:

2     the parsing module structuring each such identified term in the one or

3     more concepts into canonical concepts comprising at least one of word root,

4     character case, and word ordering.

1     27.     (original): A system according to Claim 25, wherein at least one of

2     nouns, proper nouns and adjectives are included as terms.

1     Claims 28-30 (canceled).

1     31.     (previously presented): A system according to Claim 18, further

2     comprising:

3     the similarity submodule calculating the similarity in accordance with the

4     formula:

5

$$\cos \sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A||\vec{S}_B|}$$

6     where $\cos \sigma_{AB}$ comprises a similarity between a document $A$ and a document $B$,

7     $\vec{S}_A$ comprises a score vector for document $A$, and $\vec{S}_B$ comprises a score vector for

8     document $B$.

1     Claims 32-34 (canceled).

1     35.     (currently amended): A method for providing efficient document

2     scoring of concepts within and clustering of documents in an electronically-stored

3     document set, comprising:

4     scoring a document in an electronically-stored document set, comprising:

5     determining a frequency of occurrence of at least one concept

6     within a document;

7     analyzing a concept weight reflecting a specificity of meaning for

8     the at least one concept within the document;

9     analyzing a structural weight reflecting a degree of significance

10    based on structural location within the document for the at least one concept;

11        analyzing a corpus weight inversely weighing a reference count of

12   occurrences for the at least one concept within the document; and

13        evaluating a score to be associated with the at least one concept as

14   a function of a summation of the frequency, concept weight, structural weight,

15   and corpus weight;

16        forming the score assigned to the at least one concept as a normalized

17   score vector for each such document in the electronically-stored document set;

18        determining a similarity between the normalized score vector for each

19   such document as an inner product of each normalized score vector;

20        grouping the documents by the score into a plurality of clusters,

21   comprising:

22        ~~evaluating~~ selecting a set of candidate seed documents ~~selected~~

23   from the electronically-stored document set;

24        identifying seed documents by applying the similarity ~~as a best fit~~

25   to each such candidate seed document and selecting those candidate seed

26   documents that are sufficiently unique from other candidate seed documents as

27   the seed documents;

28        identifying a plurality of non-seed documents;

29        determining the similarity between each non-seed document and a

30   center of each cluster; and

31        assigning each non-seed document to the cluster with [[the]] a best

32   fit, subject to a minimum fit; and

33        relocating outlier documents, comprising:

34        determining the similarity between each of the documents grouped

35   into each cluster based on the center of the cluster and the scores assigned to each

36   of the at least one concepts in that document;

37        dynamically determining a threshold for each cluster as a function

38   of the similarity between each of the documents; and

39        identifying and reassigning each of the documents with the

40   similarity falling outside the threshold.

1    36.    (previously presented): A method according to Claim 35, further

2    comprising:

3         evaluating the score in accordance with the formula:

4    $$S_i = \sum_{1 \to n}^{j} f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

5    where $S_i$ comprises the score, $f_{ij}$ comprises the frequency, $0 < cw_{ij} \leq 1$ comprises

6    the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$

7    comprises the corpus weight for occurrence $j$ of concept $i$.

1    37.    (previously presented): A method according to Claim 36, further

2    comprising:

3         evaluating the concept weight in accordance with the formula:

4    $$cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

5    where $cw_{ij}$ comprises the concept weight and $t_{ij}$ comprises a number of terms for

6    occurrence $j$ of each such concept $i$.

1    38.    (previously presented): A method according to Claim 36, further

2    comprising:

3         evaluating the structural weight in accordance with the formula:

4    $$sw_{ij} = \begin{cases} 1.0, & if(j \approx SUBJECT) \\ 0.8, & if(j \approx HEADING) \\ 0.7, & if(j \approx SUMMARY) \\ 0.5 & if(j \approx BODY) \\ 0.1 & if(j \approx SIGNATURE) \end{cases}$$

5    where $sw_{ij}$ comprises the structural weight for occurrence $j$ of each such concept $i$.

1    39.    (previously presented): A method according to Claim 36, further

2    comprising:

3         evaluating the corpus weight in accordance with the formula:

4
$$rw_{ij} = \begin{cases} \left(\dfrac{T - r_{ij}}{T}\right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

5  where $rw_{ij}$ comprises the corpus weight, $r_{ij}$ comprises a reference count for

6  occurrence $j$ of each such concept $i$, $T$ comprises a total number of reference

7  counts of documents in the document set, and $M$ comprises a maximum reference

8  count of documents in the document set.

1      40.     (previously presented): A method according to Claim 36, further

2  comprising:

3          compressing the score in accordance with the formula:

4          $S'_i = \log(S_i + 1)$

5  where $S'_i$ comprises the compressed score for each such concept $i$.

1      41.     (original): A method according to Claim 35, further comprising:

2          maintaining concepts and terms in a global stop concept vector cache; and

3          filtering selection of the at least one concept based on the concepts and

4  terms maintained in the global stop concept vector cache.

1      42.     (original): A method according to Claim 35, further comprising:

2          identifying terms within at least one document in the document set; and

3          combining the identified terms into one or more of the concepts.

1      43.     (original): A method according to Claim 42, further comprising:

2          structuring each such identified term in the one or more concepts into

3  canonical concepts comprising at least one of word root, character case, and word

4  ordering.

1      44.     (original): A method according to Claim 42, further comprising:

2          including as terms at least one of nouns, proper nouns and adjectives.

1     Claims 45-47 (canceled).

1     48.    (previously presented): A method according to Claim 35, further

2     comprising:

3           calculating the similarity in accordance with the formula:

4     $$\cos \sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A||\vec{S}_B|}$$

5     where $\cos \sigma_{AB}$ comprises a similarity between a document $A$ and a document $B$,

6     $\vec{S}_A$ comprises a score vector for document $A$, and $\vec{S}_B$ comprises a score vector for

7     document $B$.

1     Claims 49-51 (canceled).

1     52.    (currently amended): A computer-readable storage medium

2     holding code for providing efficient document scoring of concepts within and

3     clustering of documents in an electronically-stored document set, comprising:

4           code for scoring a document in an electronically-stored document set,

5     comprising:

6                 code for determining a frequency of occurrence of at least one

7     concept within a document;

8                 code for analyzing a concept weight reflecting a specificity of

9     meaning for the at least one concept within the document;

10                code for analyzing a structural weight reflecting a degree of

11    significance based on structural location within the document for the at least one

12    concept;

13                code for analyzing a corpus weight inversely weighing a reference

14    count of occurrences for the at least one concept within the document; and

15                code for evaluating a score to be associated with the at least one

16    concept as a function of a summation of the frequency, concept weight, structural

17    weight, and corpus weight;

18        code for forming the score assigned to the at least one concept as a

19    normalized score vector for each such document in the electronically-stored

20    document set;

21        code for determining a similarity between the normalized score vector for

22    each such document as an inner product of each normalized score vector;

23        code for grouping the documents by the score into a plurality of clusters,

24    comprising:

25        code for ~~evaluating~~ <u>selecting</u> a set of candidate seed documents

26    ~~selected~~ from the electronically-stored document set;

27        code for identifying seed documents by applying the similarity ~~as a~~

28    ~~best-fit~~ to each such candidate seed document <u>and selecting those candidate seed</u>

29    <u>documents that are sufficiently unique from other candidate seed documents as</u>

30    <u>the seed documents</u>;

31        code for identifying a plurality of non-seed documents;

32        code for determining the similarity between each non-seed

33    document and a center of each cluster; and

34        code for assigning each non-seed document to the cluster with

35    [[the]] <u>a</u> best fit, subject to a minimum fit; and

36        code for relocating outlier documents, comprising:

37        code for determining the similarity between each of the documents

38    grouped into each cluster based on the center of the cluster and the scores

39    assigned to each of the at least one concepts in that document;

40        code for dynamically determining a threshold for each cluster as a

41    function of the similarity between each of the documents; and

42        code for identifying and reassigning each of the documents with

43    the similarity falling outside the threshold.

1      53.     (currently amended): An apparatus for providing efficient

2    document scoring of concepts within and clustering of documents in an

3    electronically-stored document set, comprising:

4      means for scoring a document in an electronically-stored document set,

5   comprising:

6          means for determining a frequency of occurrence of at least one

7   concept within a document;

8          means for analyzing a concept weight reflecting a specificity of

9   meaning for the at least one concept within the document;

10         means for analyzing a structural weight reflecting a degree of

11  significance based on structural location within the document for the at least one

12  concept;

13         means for analyzing a corpus weight inversely weighing a

14  reference count of occurrences for the at least one concept within the document;

15  and

16         means for evaluating a score to be associated with the at least one

17  concept as a function of a summation of the frequency, concept weight, structural

18  weight, and corpus weight;

19      means for forming the score assigned to the at least one concept as a

20  normalized score vector for each such document in the electronically-stored

21  document set;

22      means for determining a similarity between the normalized score vector

23  for each such document as an inner product of each normalized score vector;

24      means for grouping the documents by the score into a plurality of clusters,

25  comprising:

26         means for ~~evaluating~~ selecting a set of candidate seed documents

27  ~~selected~~ from the electronically-stored document set;

28         means for identifying seed documents by applying the similarity ~~as~~

29  ~~a best fit~~ to each such candidate seed document and selecting those candidate seed

30  documents that are sufficiently unique from other candidate seed documents as

31  the seed documents;

32         means for identifying a plurality of non-seed documents;

33    means for determining the similarity between each non-seed

34  document and a center of each cluster; and

35    means for assigning each non-seed document to the cluster with

36  [[the]] a best fit, subject to a minimum fit; and

37    means for relocating outlier documents, comprising:

38    means for determining the similarity between each of the

39  documents grouped into each cluster based on the center of the cluster and the

40  scores assigned to each of the at least one concepts in that document;

41    means for dynamically determining a threshold for each cluster as

42  a function of the similarity between each of the documents; and

43    means for identifying and reassigning each of the documents with

44  the similarity falling outside the threshold.